Using computational linguistics to enhance protest event analysis

Wüest, Bruno ; Rothenhäusler, Klaus ; Hutter, Swen

Abstract: For now more than four decades, quantitative protest event analysis (PEA) has routinely contributed to the testing and refinement of theories on political processes from different perspectives. However, it is commonly agreed that PEA data face serious challenges regarding their data sources. Precisely, researchers applying PEA struggle with the fact that they cannot use multiple sources for large geographical areas and long time periods. As a consequence, most of the scholarship still focuses on a narrow set of European countries or the United States and does not cover the period since the early 2000s. We are bringing PEA and computational linguistics together to suggest and evaluate an approach that will enable political scientists to extend their research designs with a more efficient and at the same time reliable data collection. The approach relies on hidden topic models, word space models, and named entity recognition to identify and code protest events.

Using computational linguistics to enhance protest event analysis

Bruno Wüest*, Klaus Rothenhäusler†and Swen Hutter‡, November 29, 2013

Paper delivered for presentation at the ENCoRe Workshop "Tools and Techniques for Conflict Event Data Collection", Dec 6–7, 2013, University of Konstanz

Abstract

For now more than four decades, quantitative protest event analysis (PEA) has routinely contributed to the testing and refinement of theories on political processes from different perspectives. However, it is commonly agreed that PEA data face serious challenges regarding their data sources. Precisely, researchers applying PEA struggle with the fact that they cannot use multiple sources for large geographical areas and long time periods. As a consequence, most of the scholarship still focuses on a narrow set of European countries or the United States and does not cover the period since the early 2000s. We are bringing PEA and computational linguistics together to suggest and evaluate an approach that will enable political scientists to extend their research designs with a more efficient and at the same time reliable data collection. The approach relies on hidden topic models, word space models, and named entity recognition to identify and code protest events.

Contents

*Department of Political Science, University of Zurich. Email: wueest@ipz.uzh.ch
†Institute for Natural Language Processing, University of Stuttgart. Email: rothenha@gmail.com
‡Max Weber Fellow, European University Institute. Email: Swen.Hutter@EUI.eu

1 Introduction

For now more than four decades, quantitative protest event analysis (PEA) has routinely contributed to the testing and refinement of theories on political processes from different perspectives (Davenport, 2010). Most event data come from newspapers or other news archives, which is justified since news media documents are easily accessible, available for long time-series and allow a systematic coding of events that gained the broader public's attention. However, this research strategy also meets three serious challenges. First, researchers applying PEA struggle with the bias of the media outlets used. It is not fully clear how substantial differences among news media outlets in terms of partisanship, geographical coverage, and quality standards affect the results. Therefore, the use of multiple sources seems preferable but is not always done. The second issue arises with respect to the countries under study. Mainstream social movement theory is grounded in empirical cases from the US and Western Europe, and this base clearly shaped the theory. Case studies from other regions may highlight the limitations of prior theory to some extent, but differences in the methodological approaches usually impair comparability. In addition, the use of multiple sources seems even more prevalent in less stable political contexts (e.g. on the former USSR, (see Beissinger, 2002)). The final issue pertains the time periods included into the studies of protest politics, since especially the most recent years are surprisingly weakly covered.

If research applying PEA could process more than a few newspapers in a country, and if it could include more than a few countries and years into the sample, these issues would become manageable. However, the time-consuming nature of manual data collection approaches the lacking reliability of previous automation attempts. A more efficient and at the same time reliable data collection therefore constitutes a major methodological priority for protest researchers. We are bringing PEA and computational linguistics together to suggest such an approach and present results from an in-depth comparison with manual PEA data. The approach relies on hidden topic models, word space models, and named

entity recognition to select relevant news documents and to identify the concepts of a protest event. On the basis of our previous experience with electoral campaigns (Wueest et al., 2011), we understand our task insofar as we try to establish semi-automatic procedures to collect a basic set of key indicators of protest events such as the protest form, the number of participants and the location. The specific classifications of these variables as well as any additional variables should be left to specific PEA projects.

This paper is structured in three parts. The first part will discuss the current state of the art in quantitative protest event analysis. As already outlined, the existing approaches suffer from major limitations regarding their feasibility for large-n comparisons involving many country and news sources. The second part introduces the technical implementation of our semi-automated approach to select and code protest events. We harness two recently introduced computational linguistic methods, namely hidden topic and word space models, for content analysis in the political sciences. The final part will then evaluate our semi-automated content analysis against a data set established in an actual study on political protests in Western Europe.

Our main motivation to look for possible computational enhancements to PEA content analyses stems from the need to tackle the limitations of the manual applications we did so far (e.g., see Kriesi et al., 2012). The paper, however, is in its early stages and might still give the impression of a workbench note. Most obviously, it is work in progress as some implementations and evaluations regarding the coding of protest events are not established yet. Yet the existing results so far point to the feasibility of our approach also for other studies on social movements and political contention. Furthermore, we present some computational linguistic techniques which are new to the political sciences in general.

2 Protest event analysis: an overview

This section provides a brief overview of the use of PEA in social movement research. We underscore the centrality of the method to the field and its usefulness in answering a variety of research questions. Furthermore, we emphasize that social movement scholars have paid close attention to the selection bias of PEA data. However, due to resource constraints and efficiency considerations, scholars still often rely on a limited number of sources only. In addition, the geographical scope of most data sets is restricted to a few Western industrial countries, and only a few data sets cover the period since the early 2000s (for more detailed accounts, see Koopmans and Rucht, 2002; Davenport, 2010, 25ff.).

Researchers rely on PEA, as a type of content analysis, to systematically assess the amount and features of protests across various geographical areas (from the local level up to the supranational level) and over time (from short periods of time up to several decades). Usually, social movement scholars use newspapers articles as their text sources, but the range of sources has expanded over time and covers, amongst others, police reports and information provided by new digital media. To move beyond a few cases and illustrative examples is also what made PEA so attractive to social movement scholars. As Koopmans and Rucht (2002, 252) have aptly stated, "PEA provides a solid ground in an area that is still often marked more by more or less informed speculation."

Since early work in the 1960s and 1970s, this has led to a "a virtual industry of protest event data analysis" (Klandermans and Staggenborg, 2002, xii). (Oliver, Cadena-Roa and Strawn, 2003, 214) list the increasing use of PEA even among the top-four emerging trends in social movement research, which they believe to be important in the coming years and that "all involve transcending old categories and boundaries and all combine methodological and theoretical advances." Since the beginning in the late 1960s, one can identify at least four generations of PEA research, and the main 'boom' period were clearly the late 1980s and 1990s (see Crist and McCarthy, 1996; Koopmans and Rucht,

2002, 232ff.).

The first generation consisted of researchers who were interested in various indicators for a large number of countries or in long-term processes of social and political change. The Handbook for Social and Political Indicators I & II by Russett et al. (1967) as well as by Taylor and Hudson (1972) are the most prominent examples for large N-studies. Tilly and his colleagues, by contrast, were interested in the long-term trends of strike activity and political violence (Shorter and Tilly, 1974). However, these authors paid relatively little attention to "the selectivity of the sources, the creation of fine-grained coding categories, and the development of well-documented rules and procedures" (Koopmans and Rucht, 2002, 232). This led to first methodological debates over the selectivity of newspaper reports (see the interesting debate between Danzger (1975) and Snyder and Kelly (1977)).

Inspired by this research, a second generation developed that made more extensive use of protest data. This research broke down the data according to various analytical criteria, which was possible as the categories used for the data collection were far more sophisticated. Path breaking studies were Jenkins and Perrow's (1977) work on farmers' mobilization in the United States, Kriesi et al.'s (1981) study on political activation events in Switzerland, McAdam's (1982) case study on civil rights protests in the United States, and Tarrow's (1989) study on the Italian protest cycle from 1965 to 1974. These studies mainly focused on the emergence and development of social movements due to 'expanding opportunities'.[1] Furthermore, a major innovation within this second generation were cross-national designs, such as that used by Kriesi et al. (1995) to study new social movements in four West European countries. Here, the focus was more on the more stable elements of the political context faced by social movements and how this context affects the levels and forms of mobilization (on environmental protests, see Rootes (2003)).

Though the second generation was sophisticated with respect to coding procedures

[1] McAdam's (1982) well-known study, for example, traced the development of the civil rights movement in the United States by focusing on changes in demography, repression, and the political economy (e.g., the collapse of the cotton economy).

and source selection, the authors did not invest a lot of time in qualifying the bias of their sources. Thus, a third generation assessed the bias of newspaper data more systematically. Overall, researchers still disagree on how bad the selection bias is. For example, in recent review articles, Earl et al. (2004) draw a rather optimistic conclusion, while Ortiz et al. (2005) emphasize the shortcomings of PEA. Earl et al. (2004: 69ff.) stress three sets of factors that predict selection bias and increase the news value of a given protest event:

(1) Event characteristics: Many studies have shown that large and violent events are more likely to be reported than small and peaceful ones (e.g. Barranco and Wisler, 1999; Fillieule, 1996; Hocke, 2002; McCarthy, Smith and Zald, 1996; McCarthy et al., 2008; Oliver and Maney, 2000; Oliver and Myers, 1999). Another event characteristic increasing coverage rates is the proximity to the news agency–be it internationally (e.g. Mueller, 1997) or regionally (e.g. Davenport, 2010; Snyder and Kelly, 1977). Other event characteristics, which increase coverage rates, refer to the presence of counterdemonstrators or police forces, sponsorship by formal organizations, and the significance of the actors involved (e.g. Hocke, 2002; McCarthy et al., 2008; Myers and Schaefer Caniglia, 2004; Oliver and Maney, 2000);

(2) News agency characteristics: Danzger (1975) showed years ago that the presence of a wire service in a city increases the likelihood that an event is being covered. Oliver and Myers (1999) show, for example, that 'routinized' events confirming to expectations about when, how, and where events are taking place are more likely to be covered by journalists than 'non-routinized' events. Additional variables refer to audience characteristics and the self-definition of newspapers.[2]

(3) Issue characteristics: Protest events, which resonate with more general concerns, are more likely to be reported. McCarthy et al.'s (1996) study on Washington,

[2]For example, local newspapers are less selective than national newspapers (Swank, 2000). Liberal or extreme left newspapers are less selective than conservative papers (e.g. Koopmans, 1995; Oliver and Myers, 1999). Similarly, Davenport's (2010) case study on the Black Panther Party shows that sources closer either to authorities or dissidents are more likely to cover events, while sources somewhere between the two extremes are less attentive.

D.C., is most often cited as showing such effects of 'media attention cycles'. In another local study, Oliver and Maney (2000) showed that legislative conflict over a particular issue raises the likelihood of protest being covered. Overall, the results on issue characteristics are less clear-cut than on event and news agency characteristics (Ortiz et al., 2005, 401), but the diverse range of studies cited shows that the results are rather consistent across different countries and media sources (McCarthy et al., 2008, 142).

Apart from assessing the selection bias of newspaper data, part of the third PEA generation tried to be more efficient by using electronic approaches to select (and even code) protest events. Most prominent examples of half-automated procedures are:

(a) the 'European protest and coercion data' (EPCD) collected by Francisco et al. (e.g. Francisco, 1996; Nam, 2006; Reising, 1999),

(b) Imig and Tarrow's (2001) study on European protest events, and

(c) Jenkins et al.'s project for a new edition of the Handbook for Social and Political Indicators.

All these projects are based on adapted version of the KEDS, the Kansas Event Data System software, to identify relevant protest events. For example, Imig and Tarrow use headlines of the Reuters news wire to identify protests motivated by EU institutions and policies from 1984 to 1997. Unfortunately, these projects tend to fall back to the first generation of research when it comes to the selection of sources and coding procedures and/or their value for comparative research (Imig 2001: 256f.). In a less ambitions way, others have used key word searches in electronic archives to reduce the time needed for the selection of relevant articles (e.g. Kriesi et al., 2012; Strawn, 2010).

Finally, a fourth generation has developed since the mid-1990s. Authors have moved beyond PEA by abandoning the strict focus on (aggregates of) protest events as their unit of analysis. On the one hand, scholars unpack single protest events and focus on

different activities and dynamic interactions within a single event (e.g. McPhail and Schweingruber, 1998; Tilly, 2008). On the other hand, scholars broaden the unit of analysis beyond protest by relying on other forms of content analysis to cover both protest events and other sorts of (mainly verbal) claims making (e.g. Koopmans and Statham, 2010; Koopmans et al., 2005). By shifting to the study of mass-mediated public debates, these scholars partly evade the question of using selective sources since they are not as interested in how biased the media sources are but in the social construction provided by the media per se.

To sum up, this brief history of PEA research has highlighted how important the technique has become in the social movement field and that researchers have invested a lot in studying the selection bias of their data sources. Thus, they paid close attention to what (Tilly, 2002, 249) has described as "event catalogues as theories". In other words, research focused on both "a theory embodying explanation of the phenomenon under investigation, and another theory embodying explanations of the evidence concerning that phenomenon" (see also Davenport, 2010; Tilly, 2008). However, the use of multiple sources in projects that cover large geographical areas and/or long time periods is still rare because of the costs and time needed to collect the data. Similarly, most of the research cited above still focuses on a selected number of (West) European countries or the United States, and do not cover the period since the early 2000s (for a summary, see Table 1). This might not only produce specific results when it comes to the selection bias of (media) sources (Strawn, 2008) but it might also affect the results on the first type of theory mentioned by Tilly. Efficiency gains in the coding procedure thus lie at the heart of potential remedies to the weaknesses of existing PEA applications.

Table 1: List of protest event data sets

Name	Geographical focus	Time period	Sources
Prodat (Rucht, 2001)	Germany	1950-2002	Two national newspapers
Political Activation in Switzerland (Kriesi et al., 1981)	Switzerland	1945-1979	Several newspapers & other sources
National political change in globalizing world (Kriesi et al., 2012)	Europe: A, CH, D, F, NL & UK	1975-2005	One national newspaper per country
European Protest and Coercion data (Francisco, 1996; Nam, 2006)	29 European & 4 Latin American countries	1980-1995	Reuters plus additional national newspapers
Europrotest I (Imig and Tarrow, 2001)	15 EU member states	1984-1997	Reuters news wires
Europrotest II(Uba and Uggla, 2011)	27 EU member states	1992-2007	News wires and national newspapers
New social movements in Western Europe (Kriesi et al., 1995)	Europe: CH, D, F & NL	1975-1989	One national newspaper per country
Demonstrations in France (Fillieule, 1997)	Marseille and Nantes	1979-1991	Local newspapers and police archives
Transformations of environmental activism (TEA) Rootes (2003)	Europe: D, F, GR, I, E, S, UK & Basque country	1988-1997	One national newspaper per country
Eastern Europe (Ekiert and Kubik, 1998)	Europe: D-East, H, SLO, & PL	1993-1998	One newspaper per country
Contentious politics in France and Britain Tilly (1986, 1995)	France and Britain	17th/18th century	Several newspapers & other sources
Protest in East Germany (Mueller, 1997)	East Germany	1989	Foreign newspapers
Protest in early modern Japan (White, 1995)	Japan	1590-1740	Several sources
Protest in Mexico (Strawn, 2008)	Mexico	2006	Local newspapers and national news wire
Former USSR (Beissinger, 2002)	Former USSR	1987-1992	Several news sources
Eastern Europe Beissinger and Sasse (2011)	11 Eastern European countries	2007-2010	Several news wires
Media Coverage of Message Events Oliver and Myers (1999)	Madison, Wisconsin	1994; 1993-96	Police archives, two local newspapers
Dynamics of Collective Action McCarthy, Rafail and Gromis (forthcoming); Soule and Earl (2005)	USA	1960-95	One national newspaper

3 Bringing in computational linguistics

This section presents the technical details of the computational linguistic software framework which supports the semi-automatic coding of protest events, thereby increasing the efficiency of PEA. Reframed as a computational linguistic task, the selection of texts covering protest events can be understood as a binary classification task which assigns to any given input text a label which denotes whether the text is concerned with a protest event or not. So what we are presenting is basically a text classification system. Furthermore, many of the techniques used to build the classifier can later be invoked for the coding of protest events.

Research in automatic text classification using machine learning techniques is abundant (for an overview, see Jurafsky and Martin, 2008; Manning and Schütze, 2002). However, there is no general best solution but instead careful feature and parameter selection are indispensable to build a well functioning text classification system. Hence, our main focus in this technical part of the paper is on feature selection and the learner we chose for the particular task of protest event identification. In text classification input texts are usually represented as a feature vector where the dimensions of the vector correspond to certain characteristics (features) that are extracted from the input and weighted in some way. Most important among these are definitely word frequencies. Using various sorts of linguistic preprocessing, however, may improve performance.

Linguistic processing

We build our system using the UIMA framework[3] which offers infrastructure for building highly modular and therefore flexible processing pipelines for textual data. The framework provides a programmatic interface for analysis modules which are implemented in the Java programming language. In a pre-processing step the textual content of the original HTML documents is extracted into a simple XML-format retaining the original

[3]http://uima.apache.org/

markup of paragraphs. The pipeline is then fed with these transformed documents and carries out the following linguistic analyses:

Tokenization is the process of identifying individual words (tokens) within a document.

Stopping refers to a technique commonly used in information retrieval where non content bearing words, so called stop words, such as articles, pronouns, etc. are removed.

Lemmatization determines the base form (lemma) of a given token, e.g. for the verb forms *protest*, *protests* and *protested* the corresponding infinitive *protest* is identified.

POS tagging (part of speech tagging) tries to find the word class for an input token. Thus, the word *protest* might be assigned to either NN (a singular proper noun) or VBP (a present tense form of a verb) depending on the context in which it occurs. The labels are taken from the Penn Treebank tag set[4] a standard for the English language. Both lemmatization and part of speech tagging are carried out using the TreeTagger[5] a widely used resource in the computational linguistics community. Whereas we use lemma information directly as input for the classification algorithm the word class is only important as an input for the next two analysis modules.

A **sentence splitter** cuts up a document into individual sentences and makes use of punctuation tags assigned by the POS tagger.

Finally in **parsing** a sentence is assigned a syntactic structure. There are different schools of syntax that advocate different kinds of syntactic analyses. We use the mate parser[6] that produces so called dependency parses. A dependency parse is a directed graph with labelled edges where the individual words are the nodes. Every node (word) is connected to exactly one other by an incoming edge. The edges are

[4]http://www.cis.upenn.edu/~treebank/
[5]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[6]http://code.google.com/p/mate-tools/

meant to indicate the relation between a head word and its modifiers (dependents). They are labelled with the grammatical relation that holds between the two.

For an example see figure 3. In the further development of our system we will try to use the information provided by the parse trees to identify certain variables of protest events such as the cause which is indicated by the solid black parts in the graphs (in this case *cuts*). At the moment though we just use the frequency of dependency triples which are combinations of a dependent, the label of its incoming edge and its head as an alternative to token or lemma frequency.

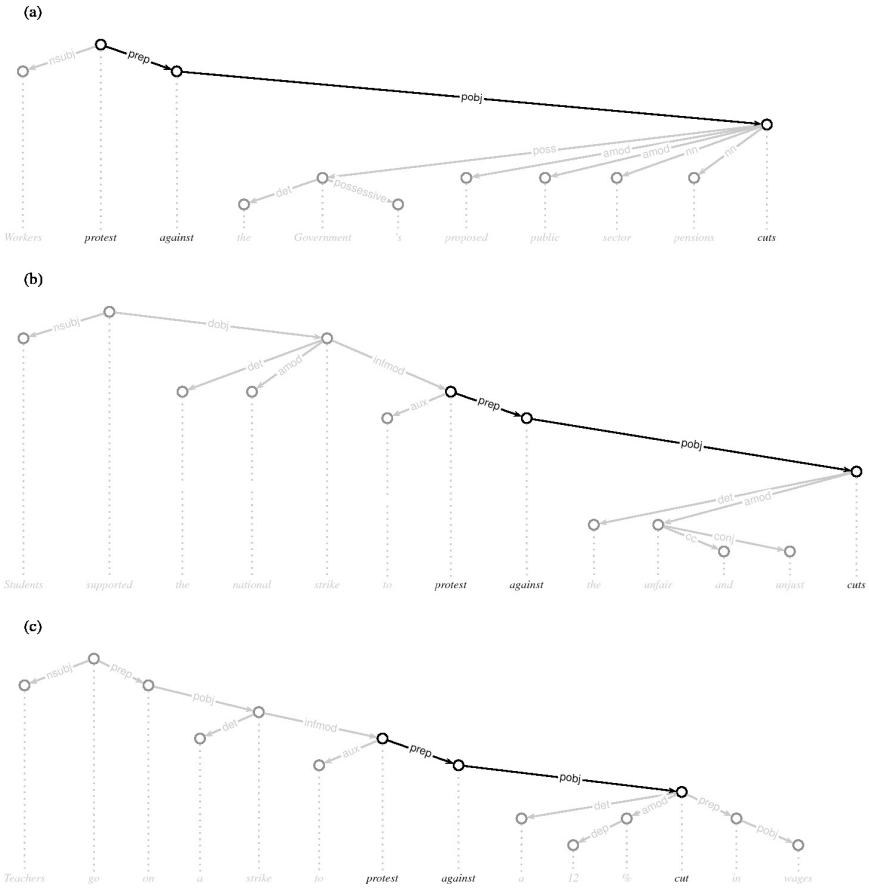

Figure 1: Automatic dependency parses

Dealing with Data Sparseness: Hidden Topic Models

Computing feature weights for frequencies counted within a document easily suffers from data sparseness problems which becomes even more aggravated when sub-textual contexts such as paragraphs or sentences are used. In order to overcome such problems we closely follow the approach proposed by Phan, Nguyen and Horiguchi (2008). Their basic idea is to build a hidden topic model from a large unlabeled set of documents and then infer the topic distributions for a given input text to compute additional more robust features to feed the classifier. Hidden topic models assume that documents are generated from a rather small number of topics which, in turn, determine the distribution over words. The model for a predefined number of topics can be computed in an unsupervised fashion using standard statistical techniques such as Gibbs Sampling. We make use of the Mallet Toolkit[7] to find the topics and produce an inference for given input texts. We worked with standard settings (1000 iterations for Gibbs sampling with a burn-in of 200 iterations for training and 100 iterations for inferences) and tested varying numbers of topics. As in the original article we found that classification accuracy is quite robust with respect to the number of topics as long as it is chosen within some reasonable range between 20 and 100. We therefore fixed the number of topics at 50 for all the experimental runs reported here.

To create a suitable topic model the unlabeled data for training must closely reflect the characteristics of the documents to be classified. In order to achieve this we again performed a keyword search using exactly the same query as above but including all the available large British quality newspapers[8]. This resulted in a total of over 21'000 documents containing over 13 million words. Different topic models were then computed by using tokens directly or in lemmatize form as well as by applying stopping or not.

[7] http://mallet.cs.umass.edu/

[8] The Independent, Daily Telegraph, Guardian, Observer, and Herald were available, which means that only one large quality newspaper, the London Times, is missing.

Choosing a learner

We use the gold standard as the labelled input for training a classifier. For that purpose we first of all split the data into a training set (70% of the data), a development set (20%) and a test set (10%). The training set is the one actually used to train a classifier model. At the current stage we monitor performance on the development set to find the best set of parameters to use. After parameter tuning is complete a single model will be selected based on the performance over the development set and run over the test set exactly once to produce performance figures for publication. Here we only present evaluation results over the development set but as the data contained in it must be considered unknown to a model built with the training set we are confident that the results will carry over well to the test set.

The overall purpose of our classifier is to reduce the amount of manual work needed for the selection of protest event descriptions. Using all the available training data as input for building the model would be equivalent to labelling 70% of the documents retrieved by the keyword search. While this reduction would be worthwhile we believe you can do better, much better indeed, by employing *active learning* to train the classifier. Active learning is a special algorithm used for classifier building where the learner itself chooses which is the most valuable example to label next. Value here corresponds to the expected performance gain for further classification that can be drawn from an example and can be computed in several ways.

In an actually implemented active learning system coders will be presented with the examples the learner chose from the set of documents collected via keyword search and asked to label them. For our experiments, however, we use the active learning implementation provided by Vowpal Wabbit[9] which allows to simulate the actual learning process given a labelled data set. The number of examples eventually used in training is controlled by a number of parameters which we do not cover here instead we just give the amount of examples used. For details concerning the active learning algorithm used by

[9]https://github.com/JohnLangford/vowpal_wabbit/wiki

the toolkit see Beygelzimer, Dasgupta and Langford (2009).

Adding named entity recognition for the protest event coding

After having a sample of relevant documents reporting on protest events, we obviously need to code the variables related to a protest event. The original codebook lists 26 variables, many of them fitted with several categories. While it is certainly necessary to adapt the PEA coding procedure to the specific requirements of a research project, we maintain that there is only a key set of variables which is relevant to all applications. Namely, virtually all PEA need to identify the protest form, issue of protest, number of participants, location and date of protest as well as participating organizations, if there is relevant information present in the document. For our approach, we will therefore focus on this restricted set of the six most generally useful variables. Yet we can also easily anticipate potential modifications of these variables or possible additional variables. Therefore, we explicitly recommend that the single research teams adapt the existing and add their new variables to this basic data set.

The coding of the number of participants, participating organizations, location, and event date requires no or only minimal interpretation by human coders and can thus be resolved by named entity recognition (NER). NER labels sequences of words in a text which are the names of entities such as persons, organizations or city names. For example, it is very straightforward to derive from the information that a protest happened 'today' that the protest actually was staged on the publication date of the newspaper article. For these tasks, we plan to embed the Stanford named entity recognizer into the software pipeline (Finkel, Grenager and Manning, 2005). Built on a conditional random field sequence model, the Stanford NER runs slower but is more precise than usual recognitions based on hidden markov models (Lafferty, McCallum and Pereira, 2001). Among others, the named entity recognizer is trained to identify persons, locations, organizations, time, and dates in British and American newswires, tasks which are very similar to recognize the following entities of protest events: organizations and persons involved in the protests,

Table 2: Classification of issues and protest forms

Issue	Form
economic	petitions, letter-writing campaigns etc.
cultural	public demonstrations etc.
security/peace	strikes/boycotts/occupations
institutional/campaign	attacks/blockades

Table 3: PEA example

Last week 50 fishermen blockaded Peterhead port in Aberdeenshire, preventing the Faroese vessel Jupiter from offloading 1'100 tonnes of the fish to a processing plant.					
issue	form	participants	location	date	organization
economic	blockades	50	Peterhead	last week	none

the locations of protest, and the day(s) when the protests were staged. The locations can be identified by labeling geographical names. Protest dates can be coded using temporals, weekdays and calendar dates. Organizations and persons, finally, can be tagged and ex post recoded into the typology needed. The only key variable for which we have to establish our own gazetteer is the number of participants. Yet this is no exceptionally difficult task, since we can start from the numbers and numerals found in the documents.

The issue and form of protests are more complex variables and need specific recognition tools. Namely, we will code issues by measuring the correspondence of the text base to the issue-specific hidden topics. In comparison to many existing approaches, we will focus on only four categories (see table 2). Yet, although we start with a small but clearly separable set of issues, the results can easily be extended into more fine-grained categories. As for the four categories of protest forms, we also distinguish four categories with differing degrees of intensity, risk, and violence. We will identify them by establishing heuristic rules around the dependency triples generated by the word space models.

Table 3 illustrates our coding scheme with an example of a protest by Scottish fishermen as it is reported in the Guardian on August 23^{th}, 2010.

4 Feasibility tests

The gold standard is provided by the research project 'The Politicization of Europe – A comparative study of six West European countries, 1970-2010'[10], conducted at the LMU Munich, which draws on quantitative content analyses to grasp how Europe is politicized. Among others, protest events were systematically collected across six European countries for the last four decades. The research team has ample of expertise on PEA (e.g. Kriesi et al., 2012; Hutter and Giugni, 2009), thus we consider the PEA data used here to reflect the state of the art in the field.

We will train and evaluate our models for the selection and coding of protest events on the test data gathered from the Guardian in 2010. The Guardian is the source selected for the U.K. in the project. We chose to look at one year in order to be able to evaluate our approach in-depth both quantitatively as well as qualitatively. In brief, our task is to translate the coding rules as close as possible into computational procedures. Subsequently, in a machine learning process, we will evaluate and enhance the procedures on a training set until it matches the quality reached in human coding. To reduce the amount of work necessary, the coding rules of the project require that only the Monday editions are used to select protest events. Moreover, only the rubric 'home' which contains political and general news was considered. This leads to a full sample of 1'787 newspaper articles. After the application of a very rough keyword list[11], 727 articles with potentially relevant articles remained. Of these, however, only 68 actually reported on protest events. This shows that simple keyword searches are not precise enough and require extensive postselection. Of course, already this first time-consuming task renders

[10] `http://www.gsi.uni-muenchen.de/forschung/forsch_einheit/ls_verg_pol_wiss/poleu/index.html`.

[11] We downloaded the articles via the common LexisNexis Interface and the following keyword list: ((submission OR submit! AND initiative OR referendum) OR petition! OR (collect! AND signature! AND campaign!) OR protest! OR demonstrat! OR manifest! OR marche! OR marchi! OR parade OR rall! OR picket! OR (human chain) OR riot! OR affray OR (letter! I/1 campaign!) OR parade OR festival OR ceremony OR (street theatre) OR (road show) OR vigil OR (consumer OR lecture OR university OR campus OR college OR school OR pupil! OR student! AND strike!) OR boycott! OR (hunger strike!) OR blockade OR (block! AND street OR traffic OR area OR site) OR sit-in OR (sit! AND strike!) OR squatter! OR (squat! AND house OR building OR area OR property) OR mutin! OR bomb! OR firebomb! OR molotov OR graffiti OR (paint! OR colour OR fire AND assault) OR attack OR arson OR incendiar! OR (fire I/1 raising) OR (set AND ablaze) OR landmine OR sabot! OR hostage! OR assassinat! OR shot OR murdered OR killed

such manual content analyses inefficient.

We will proceed as follows. For each the selection and coding of protest events, we will first assess the quality of the manual coding by comparing the gold standard to a data set newly coded by ourselves of the same articles. Subsequently, we will present our first computational solution to semi-automate the content analysis. The evaluation will mainly rely on the following standard indicators for comparing the reliability of different coding methods (see Manning and Schütze, 2002). First, we classify coded instances into true positives, false negatives, and false positives. With regard to the identification of protest events, true positives are cases recognized correctly in both the gold standard and the test data set. As for the annotation, true positives are the cases in which the compared method agrees with the classification in the gold standard. Cases that are identified as false negatives are recognized in the gold standard but not in the test data set. False positives, by contrast, are recognized or annotated in the test data set but not in the gold standard.

From the frequencies of these categories, we can compute the recall and precision of the test data set compared to the gold standard as follows:

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$
$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

The recall indicates how often a concept found or annotated by the gold standard can also be identified or annotated via the compared method. In contrast, the precision indicates how often the compared method is correct when it recognizes or annotates a concept.

Evaluating the selection of protest events

The first step of the PEA obviously is to select the relevant news document which will later serve as the sources of the content analysis. In our gold standard, the unit of measurement is a protest event, which can be covered by one or more documents. Compared to this,

we have simplified the coding guidelines so that every report on a protest event counts as a single event, whether the same event is mentioned in other documents or not. We think this makes the coding more reliable, since the instructions how different mentions of a protest are aggregated into the unit of measurement depend on the specific research motivation.[12] As for the evaluation, we will proceed as follows. First, we determine the external reliability, i.e. the difficulty with which the existing PEA coding procedure can be transferred to other studies of protest events. This way we have a benchmark for the parameter tuning of our machine learning approach.

To assess the efficiency of the manual selection procedure, we asked an inexperienced coder who never did a selection of protest events before to protocol the time he needed to code the sample of 727 articles. This yields an estimation of the initial effort to be invested by a project team that intends to apply a PEA for the first time. Our test coder spent approximately 24 hours consulting the instructions, reading through and identifying the protest events. Thus, three workdays were put into the selection of one year of protests from one news source in one country. Assuming that this result is representative, the estimated time for a modest large-n study quickly exceeds the usual budget of Ph.D. or small research projects. For example a PEA of 2 newspapers in 10 countries over 10 years would take 4'800 hours or about 578 workdays only for the first step of the content analysis.

The reliability of the manual selection for two scenarios is shown in table 4. In the first column, we indicate the reliability of our test run on the Guardian 2010 sample. This run provided 45 true positives, but also 26 false positives and 14 false negatives. This leads to a slightly not sufficient recall of 0.78[13]. As for the precision, however, the coding run of our inexperienced coder identified 26 protest events which were assumed as irrelevant by the gold standard. Even if the recall is the crucial number for the specific

[12]For example, a research team that is interested in protests at the local level might aggregate the protest events on a much lower level than scholars who are mainly interested in the national impact of protests.

[13]Following commonly accepted guidelines, we assume a benchmark of 80 % or higher as acceptable for content analysis reliabilities (see Lombard, Snyder-Duch and Bracken, 2002).

Table 4: Reliability of manual PEA selection

	inexperienced	experienced
True positives	45	402
False positives	26	18
False positives	14	44
Recall	*0.78*	*0.96*
Precision	*0.63*	*0.90*

task of selecting articles–it is much more important to identify as much relevant articles as possible than to guarantee that every identified article is correctly chosen–, the precision of 0.63 points to the difficulty of transferring the PEA coding.

In the end, the application of coding rules by humans will always involve the subjective interpretation and thus potential sources of error. A major misunderstanding of the coding instructions during our test coding actually was concerned with Wikileak's publication of classified files in 2010 – among others, the 'war logs' on the military interventions in Afghanistan and Iraq provided by Bradley Manning. Due to the U.K.'s involvement in the military intervention and the important role of the Guardian in the publication of the files protests around Wikileak were perceived to be relevant by us but not by the gold standard, were these protests were defined as foreign and thus irrelevant. In actual research, such ambiguities and inconsistencies inherent can be dealt with by additional training and–if necessary–a recasting of the selection instructions. That this can be conducted beyond reproach is shown by the second column of table 4. The numbers stem from an internal reliability test by the project we obtained our gold standard from. This PEA is based on the Swiss quality newspaper Neue Zürcher Zeitung from 1993 to 1999 and yielded a very accurate recall of 0.96 and a similarly high precision of 0.90.

For our purposes, the results of the test run constitute the lower limit against which we will test our semi-automated recognition of protest events. If we cannot come close to the reliability of an inexperienced coder, the efficiency gain of our semi-automation is upset by too much time needed for manual controls and corrections. In the following, we

Table 5: Impact of parameter options on recall at precision-recall break-even (development set)

Parameters	Estimate	Std.Error	Pr(> |t|)
Lemmatization applied	0.025	0.010	*
Dependency triples used	0.005	0100	
Full topic model (ref=no topic model)	-0.016	0.012	
Stopped topic model (ref=no topic model)	-0.004	0.012	
Sentence level (ref=paragraph)	-0.033	0.017	*
Document level (ref=paragraph)	-0.194	0.025	***
Context of 1 sentence (ref=no context)	0.004	0.016	**
Context of 2 sentences (ref=no context)	-0.026	0.016	
Context of 3 sentences (ref=no context)	-0.036	0.016	**
All contexts merged (ref=none)	-0.022	0.013	
Contiguous contexts merged (ref=none)	-0.021	0.013	
Intercept	0.525	0.018	***
N	176		
Adjusted R-squared	0.26		
F-statistic	7.11	11/176 DF	***

Significance codes: ***=0.001, **=0.01, *=0.05

will show how we chose our optimal approach by presenting the comparison of different parameter options in a simple multivariate regression (see table 5). Overall, we run 176 different selections on the training and development sets by systematically varying the options. The selections on the training set yield recalls at the precision-recall break-even[14] ranging from 0.47 to 0.96. The mean over all runs is 0.73, which is slightly below the recall of our manual test.

The first two options relate to the level of preprocessing of the text documents. They show that lemmatization, i.e. the back transfer of all words in an article to their principal part, leads to a slight improvement of the recall. The extraction of dependency triples, by contrast, do not enhance the results. However, they will be of more use for the coding of protest events. The next two options indicate that our hidden topic model actually does not significantly add to a higher recall. However, the case for the inclusion of the

[14]This break-even is commonly applied in computational linguistics. It simply indicates that we only optimize recall until it levels with the precision. To go beyond would mean that we can only enhance our selection at the cost of too much false positives.

full topic model gets corroborated by a look at the seven runs which yielded the maximal recall of 0.96 on the training set. Four of the seven runs were achieved with the full application of the hidden topic model, which is why we included this option into our final classification.

The following five options consider the unit of analysis. It was ex ante not clear which text passage would be the most relevant to identify protest events. In some articles, especially in short news reports, the whole document is concerned with a protest event. In other articles, only one sentence is dedicated to a protest event which is only one of many stories reported on. We first tested whether the whole document, the paragraph containing the keyword used in the preselection, or the single sentence around the keyword hit constitute the relevant unit of analysis. Subsequently, we also controlled for the possibility that a window of a certain number of sentences around the tagged sentence is most relevant. A context of one sentence, for example, means that we choose three sentences including the preceding as well as the subsequent sentence for the analysis. As the results indicate the paragraph performs best. With respect to the context options, the gain achieved with the one sentence context (0.004) is clearly below the loss of the sentence level compared to the paragraph (-0.033). Overall, the best performance is thus achieved at the level of paragraph, which reflects the journalistic units of an article.

The final set of options tries to account for the fact that some documents contain several keyword hits which might influence the classification. Accordingly, we tested for the possibilities that either all contexts (parameter 'all contexts merged') or only the close-by contexts (parameter 'contiguous contexts merged') should be merged. Considering the negative effects compared to the option which includes the contexts independently from each other, both options clearly fail to improve the classification.

In the end, our best performing system achieves a recall of 78.6% and a precision of 64.7% over the development set with a standard threshold (64.3% at precision—recall break-even). During training the learner queries for the label of 35.0% of the training data. Considering that the training set consists of just 70% of the available data this means

we can build a classifier that compares favorable to human coders in its performance on unseen instances by coding a mere 24.5% of the original data. The picture becomes even brighter if we include the data of the training set in the evaluation which makes sense in a realistic setting as the manually labelled data would obviously not be thrown out. With recall at 91.5% and precision at 71.7% over the training set we achieve joint performance of 88.6% recall and 70.1% precision.

Evaluating the coding of protest events

As for the selection of protest events, we will start the evaluation of protest events with an assessment of the manual procedure. The measurement of the transferability, i.e. how difficult it is for new projects to start collecting PEA data from scratch. To this aim, we tried to replicate the gold standard coding with nothing more than the corresponding coding guidelines. Table 6 indicates the reliabilities for the issues and protest forms which we have evaluated so far.

With recalls and precisions of 0.84 or higher, our inexperienced test coder produced data for both the issues and protest forms which already meet common standards. These results were achieved without any training, wich points to a high external reliability of the PEA coding scheme. In contrast to the selection of protest events, the coding of the more complex variables related to the events does not seem to be a demanding challenge for new projects in terms of the training effort necessary. As far as the time is concerned, however, the coding of protest events is even more costly. This is not surprising since the coder needs to make intellectual decisions for six variables instead of one for the selection protest events. Overall, it took our test coder approximately 41 hours to collect all information related to the protests in the U.K. in 2010. If we approximate that to our virtual example of a PEA based on 2 news sources in 10 countries and for 10 years, we would be confronted with 8'200 work hours or 987 work days. Of course, even if our test coder worked at low efficiency, this exceeds the budget of virtually all projects in political science that do not involve an entire army of annotators.

Table 6: Reliability of manual PEA annotation

	inexperienced coder	
	issue	form
True positives	45	41
False positives	4	7
False negatives	7	8
Recall	*0.92*	*0.85*
Precision	*0.87*	*0.84*

So far, we have only implemented the software necessary to support the coding of issues and protest forms. Due to several delays, however, we are not yet able to present an evaluation of these coding. Furthermore, the labeling of protest event locations, intensities (as measured by the number of participants), dates and organization depends on the Stanford NER software which we have not implemented yet. Of course, an evaluation of all semi-automatic coding will be added in further versions.

5 Conclusion

In this paper, we suggest a new, semi-automatic approach to collect data on protest events, since existing content analysis methods fail to go beyond rather restricted research designs. Moreover, our tests have shown that the selection of protest events is more difficult to replicate by inexperienced researchers than the coding itself. Thus, besides the time-consuming manual content analysis, even more time is needed to establish a reliable selection of events.

We did implement most parts of the software necessary for our semi-automatic coding – only the NER is missing. Furthermore, we have only been able to evaluate the selection of our machine learning process so far. In future versions, we will certainly present an evaluation of the coding of protest events as well. However, already for the first step of the coding procedure, we already achieved a substantial progress, since we could select protest events by using about a quarter of the original manual effort. This alone would

enable projects adopting a similar approach to include more data sources.

References

Barranco, J. and D. Wisler. 1999. "Validity and systematicity of newspaper data in event analysis." *European Sociological Review* 15:301–322.

Beissinger, M. 2002. *Nationalist Mobilization and the Collapse of the Soviet State*. Cambridge, MA: Cambridge University Press.

Beissinger, M. and G. Sasse. 2011. *An End to Patience? The 2008 Global Financial Crisis and Political Protest in Eastern Europe*. Paper prepared for the Oxford-Princeton Conference 'Popular Reactions to the economic Crisis', Nuffield College, June 25-26, 2011.

Beygelzimer, Alina, Sanjoy Dasgupta and John Langford. 2009. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09 New York, NY, USA: ACM pp. 49–56.

Crist, J. T. and J. D. McCarthy. 1996. "'If I Had a Hammer': The Changing Methodological Repertoire of Collective Behavior and Social Movements Research." *Mobilization* 1(2):87–10.

Danzger, M. H. 1975. "Validating Conflict Data." *American Sociological Review* 40:570–584.

Davenport, C. 2010. *Media Bias, Perspective, and State Repression*. Cambridge, MA: Cambridge University Press.

Ekiert, G. and J. Kubik. 1998. "Contentious Politics in the New Democracies: East Germany, Hungary, Poland, and Slovakia, 1993-98." *World Politics* 50:547–81.

Fillieule, O. 1996. *Police Records and the National Press in France. Issues in the Methodology of Data-Collection from Newspapers*. RSCAS Working Papers 1996/25. Florence: EUI.

Fillieule, O. 1997. *Strategies de la rue: Les manifestations en France*. Paris: Presses de Sciences Po.

Finkel, J. R., T. Grenager and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI.* pp. 363–370.

Francisco, R. 1996. "Coercion and Protest: An Empirical Test in Two Democratic States." *American Journal of Political Science* 40:1179–1204.

Hocke, P. 2002. *Massenmedien und lokaler Protest. Empirische Fallstudie zur Medienselektivität in einer westdeutschen 'Bewegungshochburg'*. Wiesbaden: Westdeutscher Verlag.

Hutter, S. and M. Giugni. 2009. "Protest Politics in a Changing Political Context: Switzerland, 1975-2005." *Swiss Political Science Review* 15(3):427–461.

Imig, D. and S. Tarrow. 2001. Mapping the Europeanization of Contention: Evidence from a Quantitative Data Analysis. In *Contentious Europeans: protest and politics in an emerging polity*. Lanham: Rowman & Littlefield Publishers pp. 27–49.

Jenkins, J. C. and C. Perrow. 1977. "Insurgency of the Powerless: Farm Workers's Movements, 1946-1972." *American Sociological Review* 42:249–268.

Jurafsky, D. and J. H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Klandermans, B. and S. Staggenborg. 2002. Introduction. In *Methods of Social Movement Research*, ed. B. Klandermans and S. Staggenborg. Minneapolis, MN: University of Minnesota Press pp. ix–xx.

Koopmans, R. 1995. Appendix: The Newspaper Data. In *New Social Movements in Western Europe. A Comparative Analysis*, ed. H. Kriesi, R. Koopmans, J. W. Duyvendak and M. Giugni. Minneapolis, MN: University of Minnesota Press pp. 253–273.

Koopmans, R. and D. Rucht. 2002. Protest Event Analysis. In *Methods of Social Movement Research*, ed. B. Klandermans and S. Staggenborg. Minneapolis, MN: University of Minnesota Press pp. 231–259.

Koopmans, R. and P. Statham. 2010. *The Making of a European Public Sphere*. Cambridge, MA: Cambridge University Press.

Koopmans, R., P. Statham, M. Giugni and F. Passy. 2005. *Contested Citizenship. Immigration and Cultural Diversity in Europe*. Minneapolis, MN: University of Minnesota Press.

Kriesi, H., E. Grande, M. Dolezal, M. Helbling, S. Hutter, D. Hoeglinger and B. Wueest. 2012. *Political*

Conflict in Western Europe. Cambridge, MA: Cambridge University Press.

Kriesi, H., R. Koopmans, J. W. Duyvendak and M. Giugni. 1995. *New Social Movements in Western Europe. A Comparative Analysis*. Minneapolis, MN: University of Minnesota Press.

Kriesi, H., R. Levy, G. Ganguillet and H. Zwicky. 1981. *Politische Aktivierung in der Schweiz. 1945-1978*. Diessenhofen, CH: Rüegger.

Lafferty, J., A. McCallum and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings 18th International Conference on Machine Learning, San Francisco, CA*. pp. 282–289.

Lombard, M., J. Snyder-Duch and C. C. Bracken. 2002. "Content analysis in mass communication: Assessment and reporting of intercoder reliability." *Human Communication Research* 28:587–604.

Manning, C. and H. Schütze. 2002. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McAdam, D. 1982. *Political Process and the Development of Black Insurgency 1930-1970*. Chicago, IL: University of Chicago Press.

McCarthy, J. D., C. McPhail and J. Smith. 1996. "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991." *American Sociological Review* 61:478–499.

McCarthy, J. D., J. Smith and M. N. Zald. 1996. Accessing public, media, electoral, and governmental agendas. In *Comparative perspectives on social movements: Political opportunities, mobilizing structures, and cultural framings*, ed. D. McAdam, J. D. McCarthy and M. N. Zald. Cambridge, MA: Cambridge University Press pp. 291–311.

McCarthy, J. D., L. Titarenko, C. McPhail, P. S. Rafail and B. Augustyn. 2008. "Assessing Stability in the Patterns of Selection Bias in Newspaper Coverage of Protest during the Transition from Communism in Belarus." *Mobilization* 13:127–146.

McCarthy, J. D., P. Rafail and A. Gromis. forthcoming. Recent Trends in Public Protest in the U.S.A: The Social Movement Society Thesis Revisited. In *The Changing Dynamics of Contention*, ed. J. van Stekelenburg, C. M. Roggeband and B. Klandermans. Minneapolis, MN: University of Minnesota Press.

McPhail, C. and D. Schweingruber. 1998. Unpacking Protest Events: A Description Bias Analysis of Media Records with Systematic Observations of Collective Action - The 1995 March for Life in Washington, D.C. In *Acts of Dissent. New Developments in the Study of Protest*, ed. D. Rucht, R. Koopmans and F. Neidhardt. Berlin, DE: Edition Sigma pp. 164–195.

Mueller, C. 1997. "International Press Coverage of East German Protest Events, 1989." *American Sociological Review* 62:820–832.

Myers, D. J. and B. Schaefer Caniglia. 2004. "All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969." *American Sociological Review* 69:519–543.

Nam, T. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science & Politics* April 2006:281–287.

Oliver, P. E. and D. J. Myers. 1999. "How events enter the public sphere: Conflict, location, and sponsorship in local newspaper coverage of public events." *American Journal of Sociology* 105:38–87.

Oliver, P. E. and G. M. Maney. 2000. "Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions." *American Journal of Sociology* 106:463–505.

Oliver, P. E., J. Cadena-Roa and K. D. Strawn. 2003. Emerging Trends in the Study of Protest and Social Movements. In *Political Sociology for the 21st Century*, ed. B. A. Dobratz, T. Buzzell and L. K. Waldner. Oxford, UK: Jai Press.

Ortiz, D. G., D. J. Myers, N. E. Walls and M. D. Diaz. 2005. "Where do we stand with newspaper data?" *Mobilization* 10:397–419.

Phan, Xuan-Hieu, Le-Minh Nguyen and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*. WWW '08 New York, NY, USA: ACM pp. 91–100.

Reising, U. K. H. 1999. "United in Opposition? A Cross-National Time-Series Analysis of European Protest in Three Selected Countries, 1980-1995." *The Journal of Conflict Resolution* 43:317–342.

Rootes, C. 2003. *Environmental Protest in Western Europe*. Oxford, UK: Oxford University Press.

Rucht, D. 2001. *Protest in der Bundesrepublik Deutschland. Strukturen und Entwicklungen*. Frankfurt a. M.: Campus Verlag.

Russett, B. M., H. Alker, K. W. Deutsch and H. Lasswell. 1967. *World Handbook of Political and Social Indicators*. New Haven, CT: Yale University Press.

Shorter, E. and C. Tilly. 1974. *Strikes in France, 1830-1968*. Cambridge, MA: Cambridge University Press.

Snyder, D. and W. R. Kelly. 1977. "Conflict Intensity, Media Sensitivity and the Validity of Newspaper Data." *American Sociological Review* 42:105–123.

Soule, S. A. and J. Earl. 2005. "A movement society evaluated: Collective protest in the United States, 1960-1986." *Mobilization* 10:345–364.

Strawn, K. D. 2008. "Validity and Media-Derived Protest Event Data: Examining Relative Coverage Tendencies in Mexican News Media." *Mobilization* 13:147–164.

Strawn, K. D. 2010. "Protest Records, Data Validity, and the Mexican Media: Development and Assessment of a Keyword Search Protocol." *Social Movement Studies* 9:69–84.

Swank, E. 2000. "In Newspapers We Trust? Assessing the Credibility of News Sources that Cover Protest Campaigns." *Research in Social Movements, Conflict and Change* 22:27–52.

Tarrow, S. 1989. *Democracy and Disorder: Protest and Politics in Italy 1965-1974*. Oxford, UK: Oxford University Press.

Taylor, C. L. and M. C. Hudson. 1972. *World Handbook of Political and Social Indicators*. New Haven, CT: Yale University Press.

Tilly, C. 1986. *The Contentious French*. Cambridge, MA: Harvard University Press.

Tilly, C. 1995. *Popular Contention in Great Britain, 1758-1834*. Cambridge, MA: Harvard University Press.

Tilly, C. 2002. "Event Catalogs as theories." *Sociological Theory* 20:248–254.

Tilly, C. 2008. *Contentious Performances*. Cambridge, MA: Cambridge University Press.

Uba, K. and F. Uggla. 2011. "Protest Actions against the European Union, 1992–2007." *West European Politics* 34:384–393.

White, J. W. 1995. Cylces and Repertoires of Popular Contention in Early Modern Japan. In *Repertoires and Cyles of Collective Action*, ed. M. Traugott. Durham, NC, and London, UK: Duke University Press pp. 145–171.

Wueest, B., S. Clematide, A. Bünzli, D. Laupper and T. Frey. 2011. "Electoral Campaigns and Relation Mining: Extracting Semantic Network Data from Swiss Newspaper Articles." *Journal of Information Technology and Politics* 8(4):444–463.